



The Controlled Vocabulary Manifesto

By Frances Lightsom, Alan Allwardt, Peter Schweitzer, and Lisa Zolly

Draft Report for CDI Project CDI14-LF336

October 2015

DRAFT REPORT SUBJECT TO REVISION

Acknowledgments

Stephan Zednik and Peter Fox of Rensselaer Polytechnic Institute provided expert advice in development of the ideas in this Manifesto. This work builds on a project funded by the USGS Community for Data Integration, “Use of Controlled Vocabularies in USGS Information Applications”, including Janice Gordon, Drew Ignizio, and Dalia Varanka as collaborators, and Stephan Zednik as a consultant. Don Brown and Drew Ignizio modified the Metadata Wizard to meet the system requirements recommended by the project team. We thank Jennifer Carlino for encouraging us in pursuing the project. Important ideas were contributed by a use case review panel chaired by Dalia Varanka with members Emily Fort, Greg Gunther, Giri Palanisamy, Steve Tessler, and Roland Viger.

Contents

Acknowledgments	iii
Contents	iv
Abstract	1
Vision: USGS Data Are Easily Found.....	2
Objectives.....	5
Objective 1: Make appropriate controlled vocabularies available.....	5
Objective 2: Use controlled vocabulary terms in metadata.....	6
Objective 3: Use controlled vocabularies in catalog interfaces.....	7
Objective 4: Educate and motivate the workforce.	8
Strategies	9
Strategy for Objective 1: Make appropriate controlled vocabularies available.	9
Strategy for Objective 2: Use controlled vocabulary terms in metadata.....	12
Strategy for Objective 3: Use controlled vocabularies in catalog interfaces.	14
Strategy for Objective 4: Educate and motivate the workforce.....	15
Timing, dependencies, parallel development, and new standards	16
Action Plan	17
Status at the End of Fiscal Year 2015	17
Use Cases and Functional Requirements Analysis for Vocabulary Services	17
New Functionality in USGS Vocabulary Services	20
New Functionality in Metadata Wizard	22
Use of Vocabularies in Data Catalogs	23
Next steps.....	25

Selected References	27
---------------------------	----

Figures

Figure 1. Top page of the USGS Science Catalog offers the opportunity to filter data by a list of keywords that illustrates metadata authors' choice to use multiple near-synonyms as keywords.	3
Figure 2. Opening scene from "Keywords for Better Metadata: A Morality Play in One Act," a skit presented at the USGS Community for Data Integration Annual Workshop, May 11-14, 2015. Photo credit: Daniel Wieferich, USGS.....	9
Figure 3. Process for identifying appropriate vocabularies and making them available. The process is repeated with use cases that address the needs of different USGS Programs and Mission Areas.	12
Figure 4. Phases 1 and 2 of the project in the context of the TWC Semantic Web Methodology.....	18
Figure 5. Diagram showing the three use cases and their relationships to each other. The vocabulary server at the center of the diagram provides the services that "drive" the use cases.	19
Figure 6. Screenshot showing new functionality in Metadata Wizard that makes use of the vocabulary services.....	23

Abbreviations

CDI	Community for Data Integration (USGS)
CMGP	Coastal and Marine Geology Program (USGS)
CSDGM	Content Standard for Digital Geospatial Metadata
ISO	International Organization for Standardization
EPA	U.S. Environmental Protection Agency
FGDC	Federal Geographic Data Committee
FSPAC	Fundamental Science Practices Advisory Committee (USGS)
GCMD	Global Change Master Directory (NASA)
NASA	National Aeronautics and Space Administration
NOAA	National Oceanic and Atmospheric Administration
OME	Online Metadata Editor (USGS)
RPI	Rensselaer Polytechnic Institute
SDC	Science Data Catalog (USGS)
TWC	Tetherless World Constellation (RPI)
URL	Uniform Resource Locator
USGS	U.S. Geological Survey

The Controlled Vocabulary Manifesto

By Frances Lightsom, Alan Allwardt, Peter Schweitzer, and Lisa Zolly

Abstract

We work in a time of open data, open government, and a demand for science-based decision making in response to environmental changes, increasing risks of natural hazards, increasing demands for natural resources. We also work in a time when semantic technologies can be implemented to make USGS data more readily available to meet these challenges. Our vision is a future in which USGS data is not only available but also easily found through online catalogs that provide comprehensive and focused responses to user queries.

Use of controlled vocabularies is the key to achieving this goal. The USGS can identify controlled vocabularies that have appropriate scope and detail for describing USGS data, make these available to metadata tools and catalog interfaces, and use the vocabularies competently and consistently. First steps have been taken by a project funded by the USGS Community for Data Integration, which produced vocabulary web services that linked to a metadata tool. The next steps will reach out to additional vocabularies and metadata tools, including those maintained by other organizations, and build partnerships with additional USGS stakeholders.

Vision: USGS Data Are Easily Found

Digital data sets are valuable USGS products that are in increasing demand. In a time of climate change, environmental observations are an irreplaceable record of natural conditions at a particular time, which are critical for understanding changing earth systems and for combination with data from other sources to examine complex processes. With increasingly tight budgets, scientists re-use existing data sets when possible, and funding agencies require proposals to demonstrate that suitable data are not already available before providing funding to collect or generate new data. In a time of open government, USGS must be proactive in making available the data on which our scientific conclusions rest, for use in replicating and validating our work, and also as stand-alone products. Simply making the data available is not enough. As a public agency, we must also ensure that USGS data can be found as easily as possible, especially when people do not already know that those data exist.

In the 2013 Open Data policy, the Office of Management and Budget requires USGS to make all its data available in comprehensive, easily found data catalogs, including the USGS Science Data Catalog and Data.gov. Unfortunately, although the catalogs are easily found, the data sets are not. The data catalogs use metadata records that were written for the purpose of documenting data sets for future reuse, as required by the 1994 National Spatial Data Infrastructure policy. The dataset titles and metadata keywords were often chosen to ensure that the record could be discovered using the free-text search box of an internet search engine. This resulted in long lists of keywords, often consisting of the specific scientific jargon that the metadata author would use in looking for the data plus multiple versions of the relevant high-level ISO Topic Categories. People searching the data catalogs can look for records that include a specific term, but cannot be sure the term was used in all the relevant records, nor that it was used with the desired meaning or context. Alternatively, the catalogs allow a search to be

narrowed by choosing from a list of popular keywords, but the list is dominated by variant spellings and synonyms (fig. 1). The result is that USGS data, although available, are not easily found.

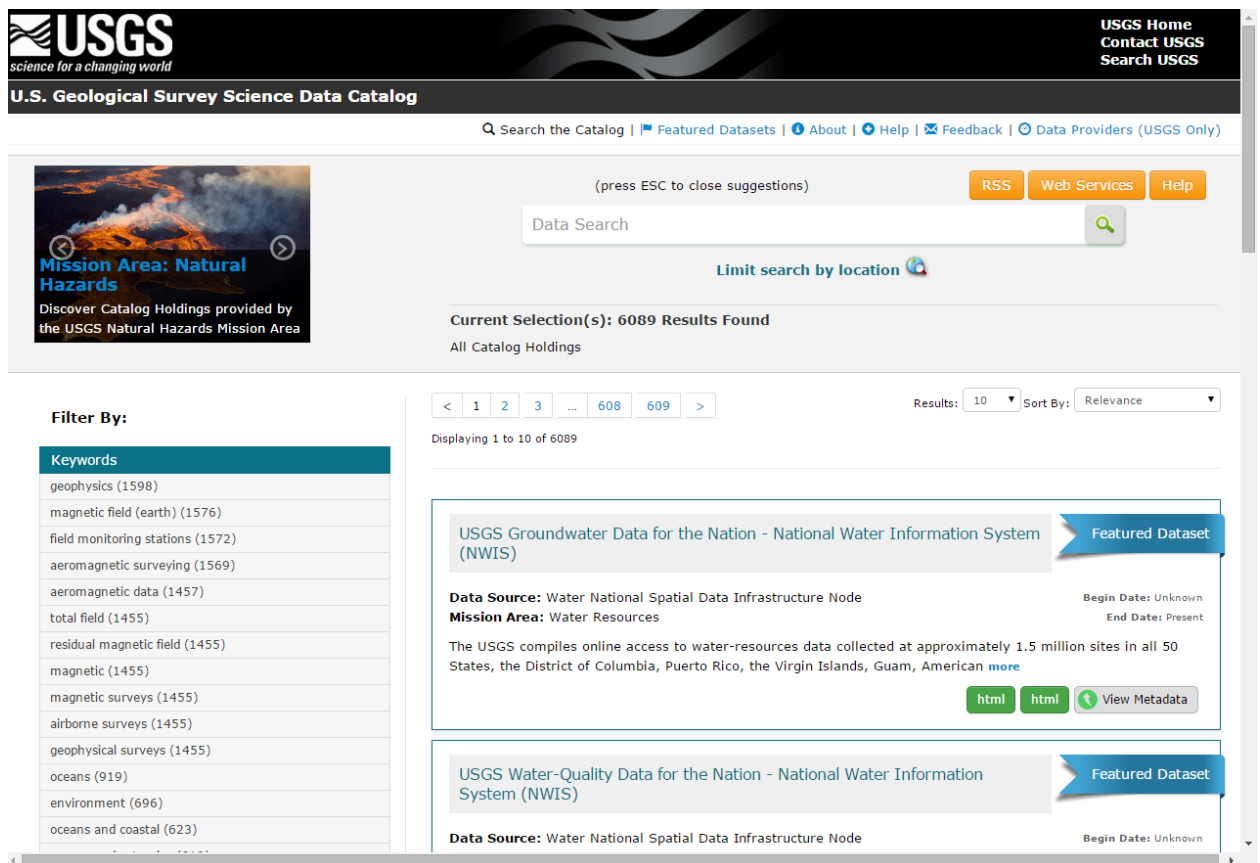


Figure 1. Top page of the USGS Science Catalog offers the opportunity to filter data by a list of keywords that illustrates metadata authors' choice to use multiple near-synonyms as keywords.

In our ideal future, people using USGS data catalogs will be confident that their search results are both comprehensive and focused, with good *recall* (nothing relevant missed) and good *precision* (nothing irrelevant included). This search will not require inside knowledge of the USGS organizational structure, the identity of USGS scientists who have worked on the topic or region of interest, or the titles of the data products. Instead, people looking for data at USGS will be offered a process that allows them to express their search criteria; the data catalog will respond with reasonable speed, providing a list of available data sets that meet the criteria, in an annotated list which can be sorted and used to evaluate

and access the data. A successful catalog will also need to provide semantic tools so that people can discover data that fits precise and nuanced criteria despite differences in terminology, definition, and context with and among the scientists who produce the data sets.

For web search engines, improving the recall of search results is generally achieved at the expense of precision, and vice versa. This conundrum is unavoidable to some degree because of the diversity of web resources: similar resources may be described in dissimilar ways, and dissimilar resources may be described in similar ways. In this chaotic information landscape the familiar free-text search box is the best option, but as a result users have become conditioned to accept search results shaped by the opposing forces of recall and precision (Brophy and Bawden, 2005). In our ideal future the users would demand more of their information retrieval systems.

Smaller, more uniform collections of information resources can offer *directed search* options that avoid some of these problems: in data catalogs like the USGS Science Data Catalog, for instance, metadata records can be mined for personal names, organizational names, topical keywords, and geographic locations, which can then be presented in browsable lists (and maps) to guide the information seeker. However, the precision and recall of such searches would be improved even more if the metadata records themselves were prepared using agreed-upon controlled vocabularies to express names, topics, and locations—and if these same controlled vocabularies were used to facilitate directed searching of the data catalog containing those metadata records. USGS Science Topics (<http://www.usgs.gov/science/>) takes this approach. The Semantic Web Working Group's 2014-15 project started developing vocabulary services that make it easier to achieve both goals: better metadata and better search interfaces. Providing comprehensive and focused catalog results for the full collection of USGS digital data will require further use of informatics techniques to provide a well-balanced set of

accommodations to technical and human limitations, consistent with the USGS mission, organization, and budget.

Objectives

Four elements will provide a framework to achieve our ideal future. Because each depends on the others, these elements are objectives to be developed in a process of linked iterations, each step building on progress toward other objectives.

1. Make available controlled vocabularies that have appropriate scope and detail for describing USGS data, for efficient and effective use throughout USGS.
2. Include controlled vocabulary terms in metadata produced by all parts of USGS.
3. Update USGS data catalog interfaces to make effective use of controlled vocabularies.
4. Educate and motivate the USGS workforce for competent and consistent use of controlled vocabularies.

Objective 1: Make appropriate controlled vocabularies available.

The use of controlled vocabularies is the keystone of the strategy. A controlled vocabulary is a set of terms that a community agrees to use for categorizing a defined set of things, with clear definitions and consistent spellings. More than one controlled vocabulary will be needed to cover the range of data products produced by the USGS. Using controlled vocabularies allows a machine to make reliable semantic matches between search criteria and metadata keywords.

Controlled vocabularies have a long history of use by libraries to assist information discovery. They improve recall by managing spelling and synonyms, so that a catalog user does not need to perform separate searches for different keywords that express the same concept. They improve precision by managing definitions and specifying the context of terms, for example, “trunk” as a part of a tree, or

an elephant, a computer network, or a car. By reducing the number of synonyms used in keyword fields, controlled vocabularies make more efficient use of interface screen space, and improve both search efficiency and catalog response speed. However, controlled vocabularies introduce additional challenges: the USGS workforce will need to learn how to use them; catalog users will experience a delay, compared to the ease of just typing in the first search term that comes to mind, and at first will not appreciate the improved quality of search results; controlled vocabularies are products and services that need maintenance.

This objective will be completed when (1) USGS identifies a set of controlled vocabularies that span the diversity of USGS data products at a useful level of detail, and (2) these vocabularies are available to the workforce and data catalogs through convenient services and tools.

Objective 2: Use controlled vocabulary terms in metadata.

For controlled vocabularies to be useful in improving the performance of USGS data catalogs, the controlled vocabulary terms need to be used in metadata records to categorize data sets. Basic metadata standards include fixed domain elements with associated code lists to enable machine interpretation of, for example, the function of a URL, or the application that can be used to display a data file. Controlled vocabularies extend this capability to metadata keyword fields, so that an automated system can match the data subject to the subject of a catalog query. But human understanding of the meaning of the terms is required to choose which keywords are appropriate for a given data set, using the definitions (or scope notes) and spellings standardized by the controlled vocabulary.

This objective builds a fundamental component of the future data catalogs that are able to offer comprehensive and focused lists of USGS data sets that meet the user's criteria. Metadata are often written by those who create and use the associated data, who will not necessarily be experts in the relevant controlled vocabularies or metadata standards. Thus, for efficiency, cost effectiveness, and to

ensure accuracy, USGS has developed tools such as Metadata Wizard and Online Metadata Editor (OME) to create metadata records, which will need to be updated to offer the capability of choosing and embedding appropriate terms from controlled vocabularies.

Work on this objective began when ISO Topic Categories were added to USGS metadata records to meet Geospatial One-Stop requirements. This objective will be completed when the metadata records that accompany new USGS data sets consistently use appropriate terms from controlled vocabularies as topic keywords.

Objective 3: Use controlled vocabularies in catalog interfaces.

As metadata using controlled vocabulary keywords become available, catalog interfaces can be modified to use them and provide better results. This will include providing people with the controlled vocabulary to use in creating their search criteria, and making use of hierarchical structures within vocabularies and mappings (crosswalks) between vocabularies to exploit the value of the controlled terms in the metadata.

Use of controlled vocabularies will allow a USGS data catalog to highlight the diversity of USGS data sets within the limited space on a catalog interface by removing the need for anyone—metadata authors, catalog users, and the catalog itself—to list all the near-synonyms that might be used to name a topic. By allowing users to choose a single top-level term in a hierarchical vocabulary like the USGS Thesaurus, a bureau-wide data catalog can offer the convenience of a smaller, topically focused catalog, while offering access to a comprehensive collection of data produced by all Programs and Mission Areas.

Geospatial One-Stop, by offering a choice among ISO Topic Categories, made the first step toward meeting this objective. When this objective is completed, (1) USGS catalog users will only need

a single search to be sure that all relevant data has been found, and (2) metadata authors will no longer need to include alternate spellings and synonyms as keywords to be sure that their data will be found.

Objective 4: Educate and motivate the workforce.

For all USGS data resources to benefit from the enhanced visibility that using controlled vocabularies can bring, it will be necessary to engage the breadth and depth of the USGS workforce in creating metadata records that include appropriate controlled vocabulary terms. A second educational effort, accustoming Americans to make use of controlled vocabularies for efficient searches through catalogs, is already underway at online shopping sites.

This objective is an opportunity for advancing compliance with open access and open data policies that require creating metadata, submitting the metadata to a USGS data catalog, and making data publicly available. The independence and variety of research centers producing USGS data will require a corresponding variety of customized approaches to achieve this objective. The diversity of research methods and kinds of data can be addressed through development of customized tools. After enthusiastic early adopters demonstrate the concept, work on this objective will benefit from respecting the expertise of data and metadata creators by including them in selecting vocabularies (for Objective 1) and designing tools (for Objective 2).

Work on this objective has begun with presentations at the USGS Community for Data Integration, including a memorable skit at the 2015 Annual Workshop (fig. 2). This objective will be completed when all new USGS metadata records include keywords from controlled vocabularies, and USGS people freely choose to add appropriate controlled terms to old metadata records that they are revising for other reasons.



Figure 2. Opening scene from “Keywords for Better Metadata: A Morality Play in One Act,” a skit presented at the USGS Community for Data Integration Annual Workshop, May 11-14, 2015. Photo credit: Daniel Wieferich, USGS.

Strategies

Strategy for Objective 1: Make appropriate controlled vocabularies available.

An appropriate set of controlled vocabularies will provide suitable scope and level of detail to enable discovery of any USGS data set, and will also meet policy requirements—for example, by using the place name list maintained by the U.S. Board on Geographic Names. If controlled vocabularies are available, they can be used by metadata tools and catalog interfaces. The strategy for meeting this objective uses recommended practices that were developed in a “discovery vocabularies” workshop convened in 2010 at Woods Hole Oceanographic Institution (Maffei, 2011). The workshop report includes guidelines for evaluating controlled vocabularies hosted in online vocabulary servers (below). Minimum requirements:

- Hosted in a stable namespace with a known technical contact.

- Serves vocabularies in a programmatically accessible format.

Additional considerations:

- Has clearly defined governance?
- Provides documentation?
- Buy-in from the target research community?
- Suitable for discovery or for usage?
- Explicitly versioned?
- Distinguishes between types and instances?
- All vocabulary terms defined?
- Extensible to new levels?
- Crosswalks to other vocabularies?

Adopting existing controlled vocabularies is preferable to developing new ones: developing and maintaining a good controlled vocabulary is time-consuming, so we should benefit from the work others have already done when possible; catalog users and metadata creators may be already familiar with the terms and definitions in existing vocabularies, which would ease implementation in USGS; controlled vocabularies are a community standard, and using interoperable standards enables data integration across the scientific community. However, there are some purposes for which no suitable vocabulary exists and for which development and maintenance of a controlled vocabulary is within the mission of the USGS as the Nation's premier earth and biological science agency. Recent cooperative work with EPA and NOAA to develop the Data Categories for Marine Planning is an example (Lightsom, Cicchetti, and Wahle, 2015).

The process for meeting Objective 1 is diagrammed in figure 3. The starting point is a use case that imagines a catalog user seeking data for a specific purpose; the use case can be analyzed to

determine the scope and level of detail needed in a vocabulary that will support that user's goal. The next step is to evaluate existing controlled vocabularies to find a good one that matches the use case. Next, the online services for the vocabulary are evaluated to see if they can interface with USGS metadata tools and data catalogs. If so, the vocabulary is added to the USGS registry of controlled vocabularies, and a new use case can be pursued. There are several branch points. If no existing vocabulary is good enough, USGS might decide to create a vocabulary (or to set aside that use case and return later when someone else might have created the needed vocabulary). If a good vocabulary exists, but is not available online with useful web services, the vocabulary would be added to the USGS vocabulary web services.

The strategy for carrying out this process is to (1) implement USGS web services that provide controlled vocabularies for our metadata tools and catalog interfaces; (2) implement a USGS registry of controlled vocabularies that identifies appropriate vocabularies, especially those that are made available by other organizations; (3) assemble a core group of use case team members with semantic web and vocabulary expertise, to which can be added subject matter experts from relevant USGS Programs and Mission Areas for the analysis of particular use cases; (4) recruit a manager to oversee the process, assist in prioritizing use cases, maintain funding, and assist in arranging cooperation with Programs and Mission Areas. The core group for use case teams would include expertise in facilitation, knowledge modeling, metadata, data catalogs, and locating and evaluating existing controlled vocabularies and their web services. Among the subject matter experts on the use case teams should be Program or science center data stewards who register metadata records in ScienceBase for inclusion in the USGS Science Data Catalog. These individuals have unique perspective on USGS data collections, and are in a position to implement the controlled vocabularies once they are chosen.

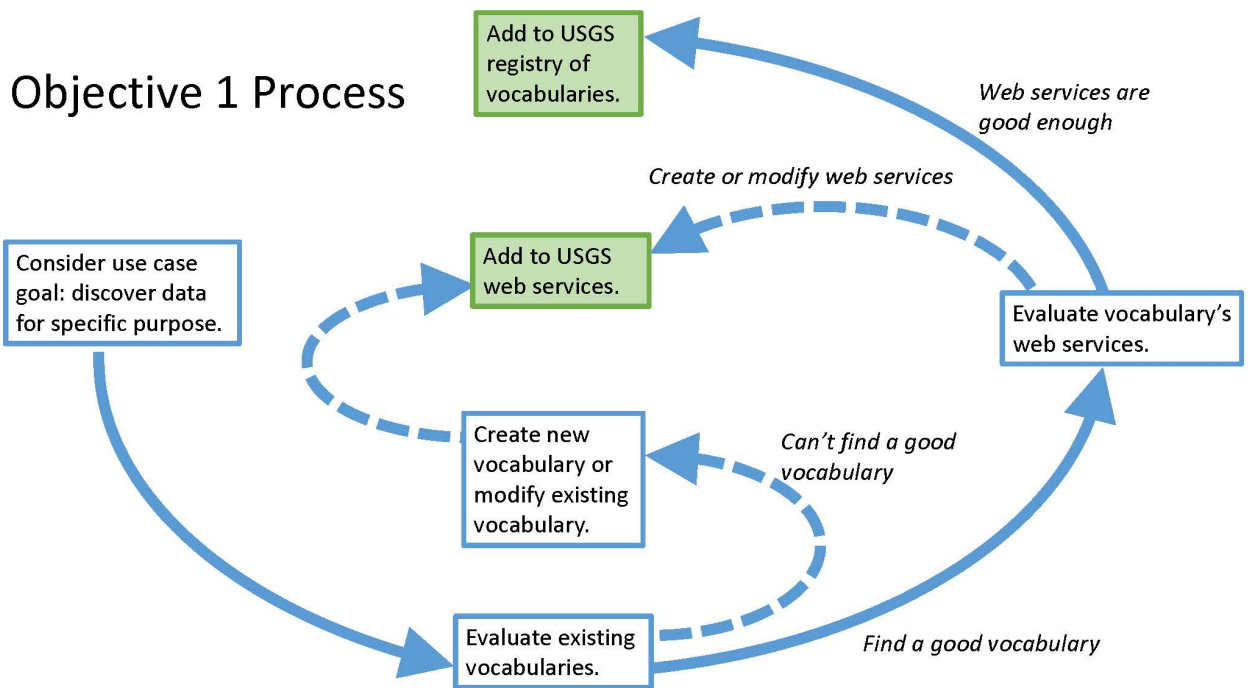


Figure 3. Process for identifying appropriate vocabularies and making them available. The process is repeated with use cases that address the needs of different USGS Programs and Mission Areas.

Depending on our success in identifying appropriate vocabularies, this objective might also require development of USGS capabilities in creation and governance of controlled vocabularies.

Strategy for Objective 2: Use controlled vocabulary terms in metadata.

The foundation for meeting Objective 2 is USGS Instructional Memorandum OSQI 2015-02 (<http://www.usgs.gov/usgs-manual/im/IM-OSQI-2015-02.html>), which requires metadata to be generated, updated, reviewed, and submitted to the USGS Science Data Catalog. Two additional efforts

are needed: modifying metadata tools to use online controlled vocabularies, and clarifying quality standards for USGS metadata to require use of controlled vocabulary keywords.

The strategy for modifying metadata tools is to work with the authors and managers of USGS metadata tools on three developments:

1. Tool modifications to make use of the controlled vocabularies provided by USGS vocabulary web services.
2. If necessary, additional tool modifications to simplify updating old metadata records, especially to replace non-controlled keyword terms with controlled terms.
3. Designing the USGS vocabulary registry so that the tools can also use appropriate controlled vocabularies whose online services are hosted by other organizations.

The strategy for clarifying metadata quality standards is to convene a community of practice for USGS metadata reviewers, including representatives from FSPAC, Bureau Approving Officials, and recognized metadata experts. This community will have many benefits and will set its own priorities, but it will be convened with the expectation that it take ownership of the USGS metadata review checklist (online at <http://www.usgs.gov/datamanagement/share/datarelease.php>) and advocate clear bureau-wide metadata quality standards. When the metadata tools have been modified to use controlled terms from the USGS web services, we will recommend that use of controlled vocabulary keywords be added to the metadata review checklist. If the community of metadata reviewers is convinced of the necessity to include controlled keyword terms, such terms will become a consistent component of new USGS metadata.

As metadata are increasingly used for data discovery as well as a source of essential information for data understanding and re-use, it will be helpful for a variety of people to make changes to metadata so that the information is effective for different purposes. For example, a scientist or technical specialist

may create the metadata, with keywords entered with the intention of indicating the detailed content of the data. But at a later stage in the process, a person managing a large collection of metadata might need to add keywords that enable similar records within that collection to be categorized similarly and consistently. USGS Fundamental Science Practices for creating and managing metadata will need to be expanded to recognize this type of flexibility in modifying existing metadata.

Strategy for Objective 3: Use controlled vocabularies in catalog interfaces.

USGS catalog interfaces already make limited use of the topic keywords in metadata, so some improvement will happen automatically with submission of a critical mass of metadata containing controlled terms. A second step is possible with the existing USGS Science Data Catalog software: providing “filter” categories that use keywords from identified controlled vocabularies. Eventually, the catalog software will be modified to make direct use of the vocabulary services, for example in up-posting (autoposting), which enables a record to be returned in response to a search for a broader topic that includes a narrower subject keyword found in the record; recommending controlled vocabulary terms to auto-complete search terms that users are entering; or using equivalent terms from different vocabularies either as suggestions or automatically.

The strategy for achieving this objective is to work with the USGS Science Data Catalog team to (1) monitor the use of controlled vocabulary terms in the metadata submitted to the Catalog, in order to recognize opportunities to make use of those terms, and (2) design interface modifications to use the improved metadata, and (3) upgrade the Catalog to interface directly with the vocabulary server to enable new data catalog functionality based on term relationships.

Strategy for Objective 4: Educate and motivate the workforce.

The strategies for the first three objectives will help educate the USGS workforce about the benefits of using controlled vocabularies, and also create incentives and penalties for additional motivation. In addition, the strategy includes development and implementation of a communication plan.

When subject matter experts from many USGS Programs and Mission Areas participate in the use case teams to identify appropriate controlled vocabularies (Objective 1), they insure that the vocabularies will be useful for their organizations, and they also carry the message back to their colleagues. By working in partnership with the teams that maintain metadata tools (Objective 2) and data catalogs (Objective 3), we will benefit from their understanding of their customer bases, and also incorporate our message into their user support processes. When we create and participate in the community of metadata reviewers (Objective 2), we will benefit from their familiarity with the diversity of USGS data and research methods, join our message with theirs, and enlist their assistance to directly require the use of controlled terms in metadata. Indirect incentives and penalties will appear when enthusiastic early adopters—so-called “positive deviants” in the organization (see Pascale and Sternin, 2005)—start using controlled vocabularies to index their records, so that these records gain visibility in data catalogs at the expense of other, less thoughtfully indexed records. This will be more effective if, when managers ask that their favorite data systems be featured, it is treated as a teachable moment rather than a demand for a work-around.

The strategy specific to Objective 4 is to convene a team to write a communication plan: identify key audiences; draft initial talking points; collect slides; identify speakers and other representatives. After the team completes its work, a communication manager will continue to maintain and update

documents and match representatives with communication opportunities. As the other strategies make progress, there will be accomplishments and positive outcomes that can be publicized.

Timing, dependencies, parallel development, and new standards

Although the strategies can be pursued in parallel after they are started, their dependencies require that major activities be phased in. Some appropriate vocabularies need to be available through the USGS web services (Objective 1) before they can be used by metadata tools (Objective 2). In turn, the metadata tools will need to be used to create a sufficient volume of records with controlled terms before the catalog interfaces can be modified (Objective 3). Attempts at general education and motivation (Objective 4) will be most effective when the other objectives are creating positive outcomes. In a different order, partnership with metadata tool development teams (Objective 2) will be needed to design the USGS registry for vocabularies that is implemented and used in Objective 1.

The topical nature of controlled vocabularies will likely result in Mission Areas and Programs being in different phases, depending on when a use case team is formed to focus on the relevant topics. However, the foundational activities within Objectives 2 and 4 can be initiated without waiting: the community of practice for USGS metadata reviewers can be convened and begin to function; the communication plan can be written.

Standards for controlled vocabularies and their online publication are not stable. Concurrent with the writing of this document, an effort is beginning within the Research Data Alliance to recommend community-based international standards for publication of controlled vocabularies on the web (<https://www.rd-alliance.org/group/vocabulary-services-interest-group/case-statement/vocabulary-services-interest-groups.html>). This is likely to result in at least one iteration of Objectives 1 & 2, as USGS technology is revised to use the new standards.

Action Plan

As this Manifesto is released, USGS has already started to achieve the objectives. Additional actions will establish a foundation for achieving the vision.

Status at the End of Fiscal Year 2015

First steps toward implementing controlled vocabularies to improve discovery of USGS data were taken as part of a project funded by the USGS Community for Data Integration during 2014 and 2015, which culminated in creation of this manifesto. The project “Use of Controlled Vocabularies in USGS Information Applications” (CDI14-LF336) developed use cases, improved existing web services for USGS controlled vocabularies, modified the Metadata Wizard tool to make use of these vocabulary services, and designed modifications in a data catalog to make use of controlled terms in metadata.

Use Cases and Functional Requirements Analysis for Vocabulary Services

The 2014-2015 CDI project was the foundation for Objective 1: Make appropriate controlled vocabularies available. The project team recognized that controlled vocabularies like the USGS Thesaurus and the Biocomplexity Thesaurus were of suitable scope and level of detail to be useful in data discovery. However, the team could not evaluate their online services. Thus the project goal was a clear definition of the services that are needed for automated use of a vocabulary in a USGS metadata tool or data catalog interface.

The project team used the TWC Semantic Web Methodology (http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology), which is designed to examine use cases and determine both functional and non-functional requirements without prejudicial commitments to meeting those requirements by utilizing particular technologies, platforms, hardware, or software. This

methodology was developed in 2008 by Peter Fox and Deborah McGuinness of the Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI), and has been taught by Fox and his collaborators to members of the project team. The TWC Semantic Web Methodology is an iterative process, as illustrated in figure 4. During the first year of the project we developed a set of use cases and a conceptual model, and engaged a panel of expert reviewers to evaluate them. Afterward, the requirements were tested by use in development of vocabulary services and modifications to a metadata tool (the Metadata Wizard) to make use of the vocabularies offered by the new services.

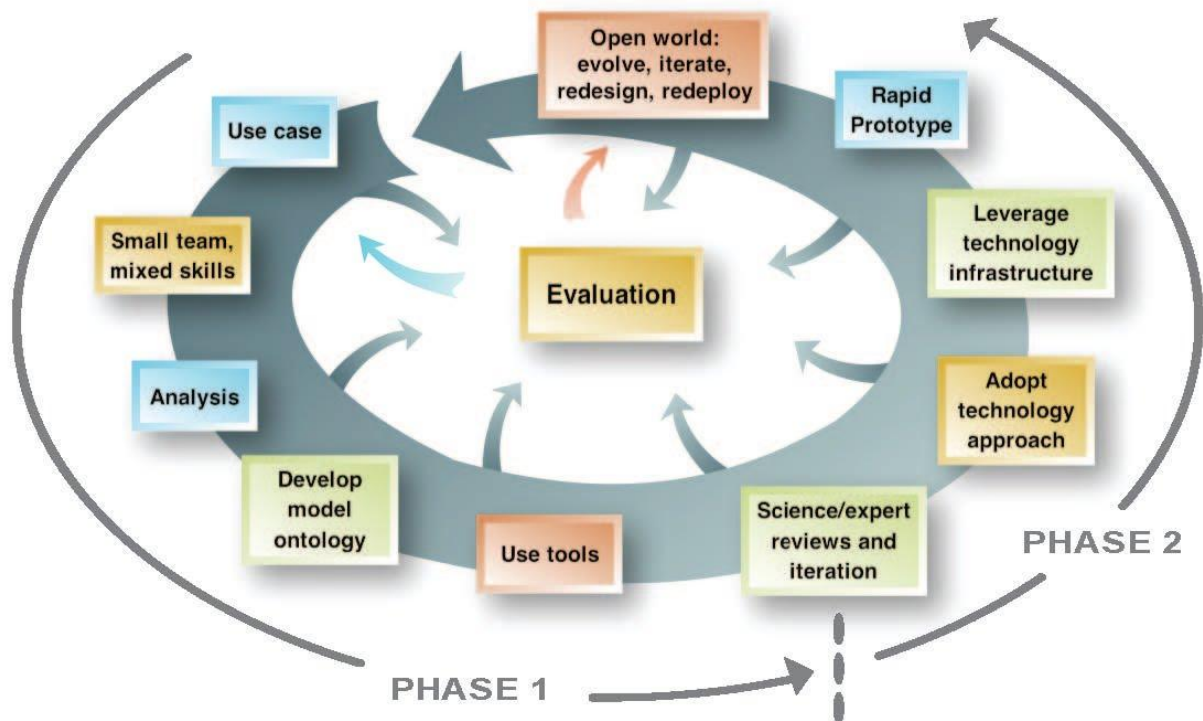


Figure 4. Phases 1 and 2 of the project in the context of the TWC Semantic Web Methodology.

The team developed three use cases for the project, as shown in the blue ovals on figure 5. The green circle at the center of the diagram represents the server that “drives” the use cases by providing the necessary vocabulary services. Use Case 1, “Assign keywords to a metadata record using one or more controlled vocabularies,” develops the functional requirements for a vocabulary server interacting

with a metadata creation tool. Use Case 2, “A catalog search interface uses vocabulary services to help users find data,” develops functional requirements for the same vocabulary server, but this time interacting with a catalog user interface. Use Case 3, “Create specialized indexes to enhance the searchability of metadata,” develops the requirements if the vocabulary server is used by a catalog system to develop an internal table that cleans up and cross-references keywords that are found in metadata records.

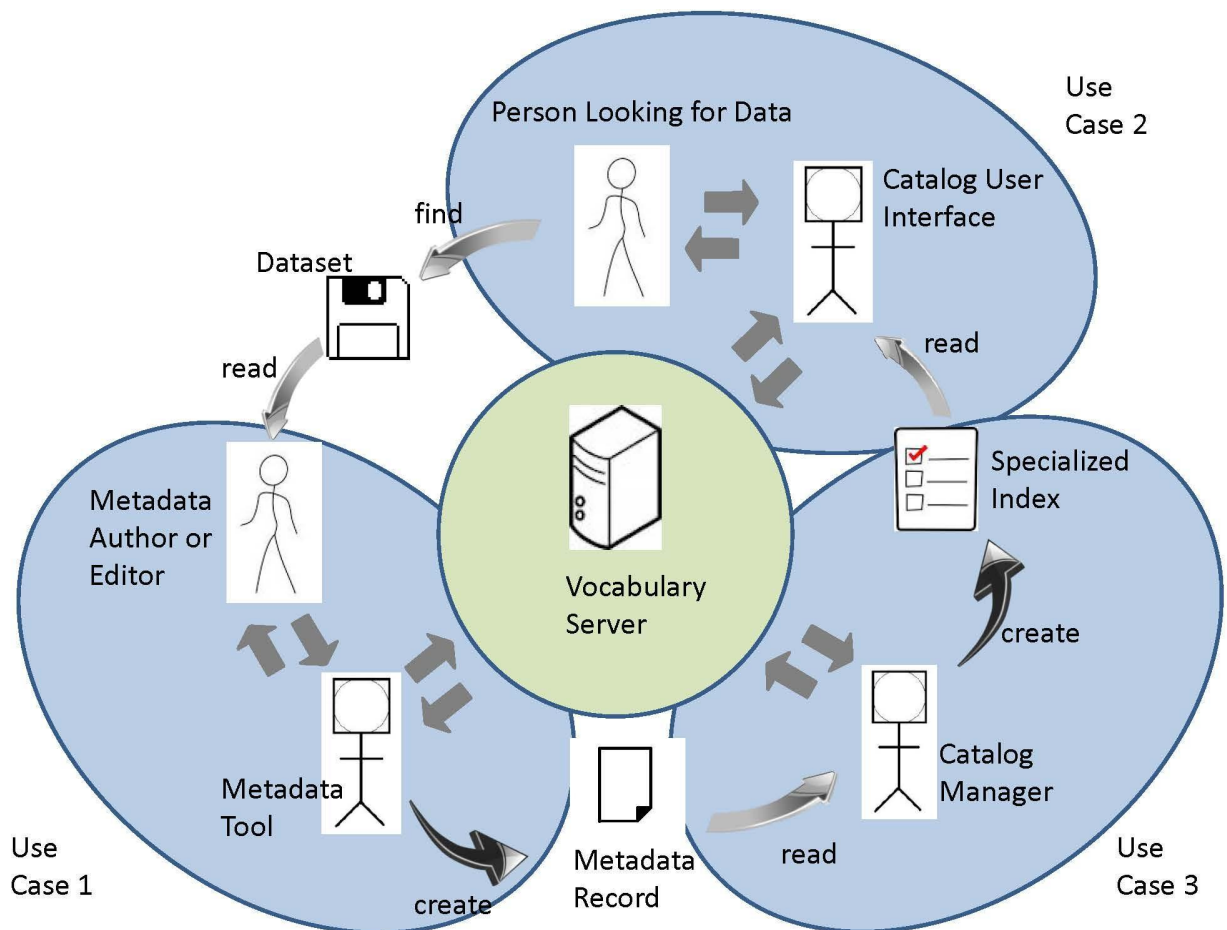


Figure 5. Diagram showing the three use cases and their relationships to each other. The vocabulary server at the center of the diagram provides the services that “drive” the use cases.

See the project website on Confluence

(<https://my.usgs.gov/confluence/display/cdi/Use+Cases+for+Vocabulary+Web+Services>) for additional

documentation of these use cases, including step-by-step narratives, visual summaries (“activity diagrams”), and conceptual models (“concept maps”). These use case documents have been revised taking into account feedback from the review panel that convened at the end of FY 2014.

Analysis of the three use cases produced a preliminary set of system requirements for a prototype vocabulary server and its web services (<https://my.usgs.gov/confluence/display/cdi/Use+Cases+for+Vocabulary+Web+Services>). In FY 2015 these requirements guided the modification of existing USGS Thesaurus vocabulary services to provide additional functionality, as described below. The requirements were also given to the Metadata Wizard developers at the USGS Fort Collins Science Center, for use in incorporating USGS Thesaurus vocabulary services in their metadata tool. More details on the Metadata Wizard work can be found on page 22.

New Functionality in USGS Vocabulary Services

Before our CDI-funded work began, the USGS Thesaurus was one of several controlled vocabularies offered through web services that were designed to meet the needs of two specific applications: the data-provisioning function of the Science Topics web interface (<http://www.usgs.gov/science/>), and parts of a prototype USGS home page designed by the Web Re-Engineering Team (<http://communities.usgs.gov/blogs/usgsdev/wret/>). The CDI project was motivated by our recognition that these specific applications had fewer functional requirements than we thought would arise in the diverse vocabulary applications the Bureau will need in the future.

Through analysis of the use cases, we discerned a need for a new service function as well as enhancements to the existing services, in order to enable applications to use multiple vocabularies. Use of multiple vocabularies will be essential to span the diversity of USGS data products at a useful level of detail. The new service function answers questions about the vocabularies that are available through

the server, and gives detailed information on each vocabulary. The enhancements enable the services to match a single search text to terms in multiple vocabularies.

To support the new service that provides information about available thesauri, the descriptive information about each thesaurus was enhanced to include a scope note describing the overall purpose of the thesaurus. As we start to use controlled vocabularies, some existing geospatial metadata use a keyword thesaurus, but do not use a controlled vocabulary term for the Theme_Keyword_Thesaurus value, so the descriptive information was also enlarged with a list of alternative names. An additional vocabulary of terms was implemented and assigned to each thesaurus to indicate some functional characteristics, such as whether the thesaurus is hierarchical or not. Additional terms were assigned to indicate the general topics addressed by each thesaurus.

The changes to the web services are summarized below. Detailed descriptions of the current web services are available at <http://www.usgs.gov/science/services.html>.

Original web services:

- term search within selected vocabulary
- term details (scope note and relationships to other terms)

Revised web services:

- list of available thesauri
- thesaurus details (alternative names, scope note, functional and topical characteristics, top-level terms),
- term search within selected vocabulary or all vocabularies
- term details (scope note and relationships to other terms)

The first step in achieving Objective 1 has been completed, and web services are currently available online. However, modifications will be required when international standards for vocabulary

services are developed. In addition, a business model and permanent web address will need to be implemented for future reliability.

New Functionality in Metadata Wizard

Metadata Wizard is a tool developed during a previous CDI-supported project to create FGDC CSDGM metadata for geospatial datasets in the Esri ArcGIS Desktop environment. Originally, Metadata Wizard allowed keywords from controlled vocabularies to be typed in. As part of the 2014-2015 CDI project, Metadata Wizard was modified to make use of the USGS vocabulary server to enable metadata authors to choose ISO Topic Categories, theme keywords from other topical vocabularies, and place keywords from place-name vocabularies (see figure 6). Other keywords can also be used, and the tool can be used to revise the keywords in existing metadata records. The prototype demonstrates vocabulary services that were identified in Use Case 1.

The new version of Metadata Wizard encourages (but does not require) at least one ISO Topic Category keyword. This is a simple list of 19 broad data themes, such as “oceans” and “biota”, so the tool retrieves the whole list from the vocabulary server and allows the user to pick which terms are appropriate.

Metadata Wizard uses the vocabulary server’s descriptive information to offer the appropriate set of vocabularies for use in theme keywords and place keywords. The user types in a search term, and matching vocabulary terms are brought back from the server, along with details about the term’s meaning, as defined by the controlled vocabulary, and related terms that are also available through the vocabulary. In addition, the descriptive information about the vocabulary is often provided, when the tool user is considering alternative vocabularies.

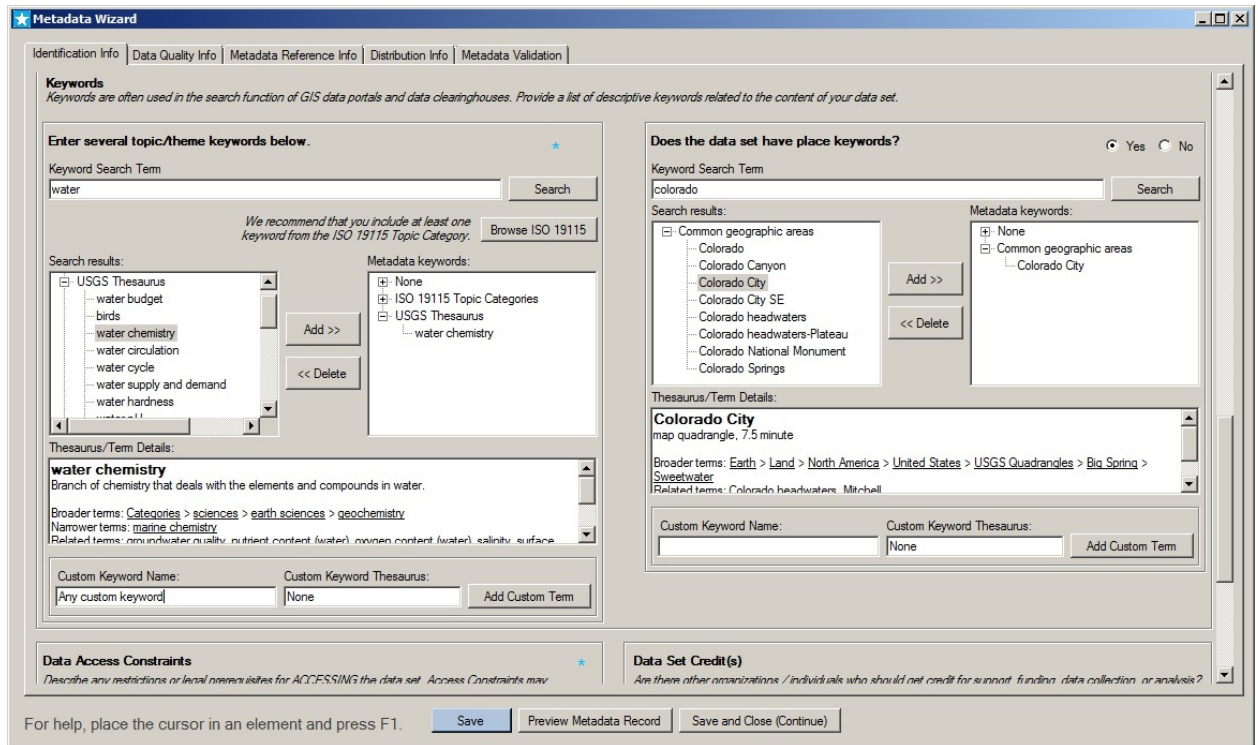


Figure 6. Screenshot showing new functionality in Metadata Wizard that makes use of the vocabulary services.

The improvements to Metadata Wizard are in version 1.7, which was released in October 2015. The new code is available through GitHub (https://github.com/dignizio-usgs/MDWizard_Source) and ScienceBase (<http://www.sciencebase.gov/metadawizard>).

Use of Vocabularies in Data Catalogs

Testing of Use Case 2 was proposed as part of the CDI-funded work, but turned out to be unrealistic because of the dependence of Objective 3 on Objectives 1 and 2. Testing of a prototype data catalog interface would require creating a sizable collection of metadata with keywords from controlled vocabularies, and then the real-world usefulness of the prototype would be limited by the extent of the metadata collection. Instead, we used the use case to make progress on two efforts: (1) our Rensselaer Polytechnic Institute (RPI) partner's development of an interface to marine planning data within

data.gov, and (2) a version of the USGS Science Data Catalog customized for Coastal and Marine Geology Program (CMGP) data.

The RPI project is applying controlled vocabularies to enable comprehensive and focused data catalog searches in support of the U.S. National Ocean Policy. The first step was development of a controlled vocabulary which is a set of categories that define the data needed for the U.S. marine planning initiative (Lightsom, Cicchetti, and Wahle, 2015), which are being used as keywords in metadata records in order to designate the associated data sets as appropriate for use in marine planning. RPI has implemented a vocabulary server for the marine planning categories, and is designing a taxonomy-based interface that will provide customized access to all data records in Data.gov that contain these terms as keywords. As of September 2015, the prototype vocabulary web services are available and the terms are being added to appropriate Data.gov metadata records so that the search interface can be tested.

Spurred by the need to update metadata records with the marine planning categories, CMGP is developing a procedure to revise large numbers of records. This is an opportunity to add additional controlled terms that will improve the discovery of the program's data. With an eye on Use Case 2, CMGP data managers are working with USGS Science Data Catalog (SDC) developers to identify low-cost modifications to the SDC system that will improve data discovery when the relevant vocabularies are used in the metadata. These modifications will be implemented in a customized user interface for CMGP data. As of September 2015, this work is proceeding deliberately as the interactions are explored between changes in the catalog interface, the metadata records, and the management of CMGP metadata. However, a first step toward Objective 3 has been achieved by developing a working relationship with the SDC team.

Next steps

The next steps will consolidate and extend the progress made by the CDI project:

- The improved vocabulary services are functionally complete but need a business model that will ensure their continuity.
- To encourage use of vocabularies that are hosted by other organizations—the NASA Global Change Master Directory, for example—a USGS registry of controlled vocabularies must be designed and implemented.
- An essential next step toward Objective 2 is implementing support for vocabulary services in additional metadata tools, starting with the Online Metadata Editor (OME) and Tkme.
- Relationships and partnerships need to be cultivated and maintained to extend the usefulness of controlled vocabularies across USGS. The USGS Publications Warehouse and the new Drupal-based USGS website are examples.

Because the USGS Thesaurus was initially developed as a method of virtually organizing USGS websites by topic rather than organizational unit, the USGS vocabulary services were created within the USGS Science Topics context (<http://www.usgs.gov/science/>). As the vocabulary services become essential to a larger range of USGS activities, a new business model is required. This action will require convening a team of stakeholders to recommend an appropriate organizational home, management structure, and stable web address (URL) that will ensure that the vocabulary server is available in the future. The team will also work with USGS managers to implement the model.

Some of the controlled vocabularies that are appropriate for describing and categorizing USGS data will be maintained and hosted by other organizations, such as the Global Change Master Directory (GCMD) maintained by NASA. To make a distributed set of vocabularies available seamlessly for automated use by tools and catalogs, USGS will need an online registry that points to the online services

for individual vocabularies that are recommended for USGS use (for example, the GCMD web services, available at <http://gcmd.gsfc.nasa.gov/Connect/>). This action will begin with an analysis of the functional requirements for the registry by a use case team that includes expertise in facilitation, knowledge modeling, metadata tools, data catalogs, as well as controlled vocabularies and their online services, using the TWC Semantic Web Methodology (http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology).

In addition to Metadata Wizard, USGS maintains other metadata tools that could be modified to use vocabulary services for producing records that have controlled keyword terms. Two of these are the Online Metadata Editor (OME, available at <https://www1.usgs.gov/csas/ome/>) and Tkme (available at <http://geology.usgs.gov/tools/metadata/tools/doc/tkme.html>). The OME is an online tool that creates metadata according to the Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), including the Biological Data Profile, and is maintained by the USGS Core Science Analytics, Synthesis, and Libraries Program. Tkme also creates FGDC CSDGM records, and is maintained by Peter Schweitzer in the Mineral Resources Program. The tool maintenance teams for both OME and Tkme will be invited to work with members of the 2014-15 project team in designing, implementing, and testing modifications. Happily, Lisa Zolly (OME) and Peter Schweitzer (Tkme) participated in the use-case development for the CDI project, so we already have a foothold for this particular “next step.”

To lay the foundations for future work, several members of the 2014-15 project team will share this Manifesto with USGS programs and offices to locate partners for future work. The Alaska Science Center and the National Climate Change and Wildlife Science Center will be approached to gauge their interest in a process of identifying appropriate controlled vocabularies and modifying metadata creation tools (Objectives 1 & 2). The Alaska Science Center, for example, has developed an ISO 19115

metadata toolkit that could utilize vocabulary services in much the same manner as the Metadata Wizard does for CSDGM metadata. We will also begin conversations with other stakeholders for the metadata reviewers' community of practice (Objective 2). Our working partnership with the USGS Science Data Catalog will be maintained (Objective 3). The USGS Communication Office will be asked for assistance in developing a Communication Plan (Objective 4)—a plan, we hope, modeled on the principle of “positive deviance” rather than “best practices.” Pascale and Sternin (2005) sum up this principle as follows: “Somewhere in your organization, groups of people are already doing things differently and better. To create lasting change, find these areas of positive deviance and fan their flames.”

Selected References

- Bittner, Kurt, and Spence, Ian, 2003, *Use case modeling*: Boston, Addison-Wesley, 347 p.
- Brophy, Jan, and Bawden, David, 2005, Is Google enough? Comparison of an internet search engine with academic library resources: *Aslib Proceedings—New Information Perspectives*, v. 57, no. 6, p. 498-512.
- Cockburn, Alistair, 2001, *Writing effective use cases*: Boston, Addison-Wesley, 270 p.
- Connaway, L.S., Dickey, T.J., and Radford, M.L., 2011, “If it is too inconvenient I’m not going after it”—Convenience as a critical factor in information-seeking behaviors: *Library & Information Science Research*, v. 33, p. 179–190.
- Fox, Peter, and McGuinness, D.L., 2008, *TWC Semantic Web Methodology*,
http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology.
- Lightsom, F.L., Cicchetti, Giancarlo, and Wahle, C.M., 2015, *Data categories for marine planning*: U.S. Geological Survey Open-File Report 2015–1046, 27 p., <http://dx.doi.org/10.3133/ofr20151046>.

- Lilly, Susan, 1999, Use case pitfalls—Top 10 problems from real projects using use cases, *in* Proceedings of the Technology of Object-Oriented Languages and Systems [TOOLS 99]: Washington, D.C., IEEE Computer Society, p. 174–183.
- Maffei, Andrew, ed., 2011, Coastal and Marine Spatial Planning Discovery Vocabularies Workshop—A joint WHOI/USGS/NOAA workshop, 12/1/2010–12/3/2010: Woods Hole, Mass., Woods Hole Oceanographic Institution, 36 p., <http://www.whoi.edu/fileservlet.do?id=77984&pt=2&p=84468>.
- Metadata Wizard (USGS), <http://www.sciencebase.gov/metadatawizard>.
- Online Metadata Editor (USGS), <https://www1.usgs.gov/csas/ome/>.
- Pascale, R.T., and Sternin, Jerry, 2005, Your company's secret change agents: Harvard Business Review, May 2005, 10 p.
- Prabha, Chandra, Connaway, L.S., Olszewski, Lawrence, and Jenkins, L.R., 2007, What is enough? Satisficing information needs: Journal of Documentation, v. 63, no. 1, p. 74–89.
- Schneider, Geri, and Winters, J.P., 2001, Applying use cases, second edition—A practical guide: Boston, Addison-Wesley, 245 p.
- USGS Thesaurus, <http://www.usgs.gov/science/about/>.